An Action-based Training Method For *VideoPose3D*

Hao Bai
ZJU-UIUC Institute
Haining, Jiaxing
https://jackgetup.com

Abstract

In this essay we perform a homogenized action-based training method based on VideoPose3D. The original paper utilizes a semi-supervised approach on different kinds of subjects and different actions in the network, and we select each action as the only input of the training dataset, and estimate the error for the corresponding action. Evidence has shown that under a limited amount of epochs and datasets, our approach outperforms the original research. This gives rise to the application for Burst Performance Instances.

1. Introduction

Vision-based human motion analysis attempts to understand the movements of the human body using computer vision and machine learning techniques. Our work focuses on the improvement of 3D human pose estimation in video. The original *VideoPose3D* model carries out the result shown in Figure 1, and our model changes the idea with Figure 2.

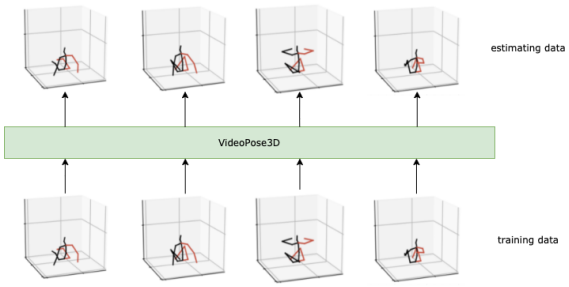

Figure 1. The original model for *VideoPose3D*

1.1. Action-based Learning

Compared to the original model we utilized a different data-processing approach. The original approach mentioned in *VideoPose3D* uses all actions in training data, and the estimating effect was good overall. We utilize the method that we only take one type of action for training, and the total amount for training should be the same. Mathematically, take the action *Sitting* for example, it can be expressed as

$$\sum_{i \in actions} f(i) = f(Sitting) \cdot t \tag{1}$$

where f stands for the frames of subject, and t means the number of epochs of our action-based training. Note that the unit of t is unit-epoch instead of epoch. The unit-epoch stands for 1 epoch of one action. According to the dataset we utilize, *Human3.6M*, there are 15 actions for each subject. Thus, the relationship between epoch and unit-epoch can be expressed mathematically as below.

$$t_0 = t_{unit} \cdot 15 \tag{2}$$

In this approach, we utilize only one time of action, which means we save more data; and we also produce the results in the same time, the evidence can be illustrated by formula (1). After testing, we conclude that our work outperforms the original work by 11%.

The principle of our work is shown below.

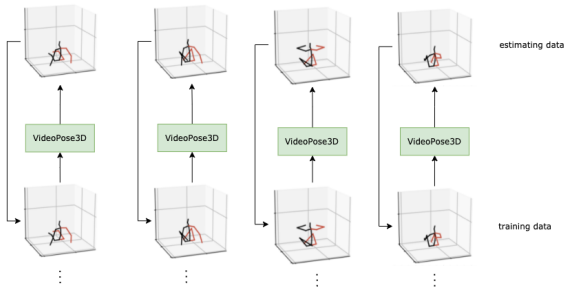

Figure 2. The action-based, multi-epoch training method for *VideoPose3D* (our work)

1.2. Iterative Training

In our model we also changed the process of training in the original research. In the original research, it takes

15 unit-epochs to run a single epoch, because there are 15 types of actions. In our work, we can tweak the number of unit-epochs not necessarily to be a multiple of 15, which makes it more flexible to observe and iterate. Shown below is the algorithm for our iterative training when not so many epochs are available (we take 2 epochs for example).

Algorithm 1: Original *VideoPose3D* Model

Data: Training data, testing data, ground-truth data, total training epoch = 2.
Result: Error value.

```
for EPOCH in 2 do
    ERR=VideoPose3D(TRAIN, TEST, GT);
    print(ERR);
end
return ERR;
```

Algorithm 2: Our Action-based Approach

Data: Training data, testing data, ground-truth data, total training epoch = 2, action type.
Result: Error value.

```
for EPOCH in 2 · 15 do
    ERR FOR ACTION=VideoPose3D(TRAIN,
      TEST, GT, ACTION);
    print(ERR FOR ACTION);
end
return ERR FOR ACTION;
```

2. Related Work

Our work is based on the previous work *VideoPose3D* with an inspiration for its semi-supervised approach. Earlier work has shown the effect of semi-supervised learning as a great way for improving training performance [8].

2.1. VideoPose3D

VideoPose3D is the current state-of-the-art approach which utilizes a fully convolutional model based on dilated temporal convolutions over 2D keypoints. [4, 5]. In early researches the main solution was to use recurrent neural networks (RNN) [2], but the research group of *VideoPose3D* utilized the convolutional neuron network (CNN) to gain an efficient result on the dataset *Human3.6M*, with the inspiration that many authors had mentioned CNN in temporal models. Some researches has also shown the idea of using a spacial-temporal model being useful [3], but that's not utilized in our research.

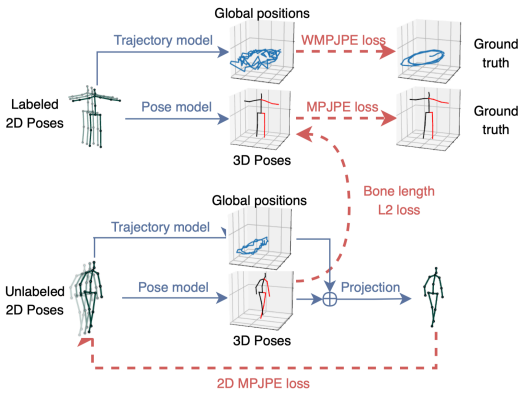

Figure 3. The overall model of *VideoPose3D*

2.2. Action-based Learning

The idea of homogenized learning was proposed firstly by Action Estimations [7, 6].

2.3. Homography Estimation

The idea of using homonegeious actions in a dataset was also inspired from Homography Estimation in some deep images [1].

3. Experimental Setup

For the code of this project, please refer to the url below.

`https://github.com/BiEchi/Pose3dDirectionalTraining`.

The detailed experimental setup can also be explored in *README.md* in the github repo. To reproduce the results, you need to test for 15 times for each type of action, and it's preferable to test for multiple times. It's more convenient to use *bash* scripts to save time.

4. Results

Shown below are the results of the experiment. Expect a minor error when testing on your own. Note that all unit are millimeter.

4.1. Small Number of Epochs

In this part we estimate the errors with a small amount of time. The epoch of the original test was set to be 1 unit epoch, and the epoch of the action-based test was set to 15 epochs.

Table 1. MPJPE, 3*3, unit-epoch

No.	Action	Original	Action-based
1	Purchases	60.55	82.09
2	Posing	57.60	70.59
3	Walking	51.82	**44.22**
4	Sitting	70.49	**70.36**
5	Walkdog	68.15	**56.14**
6	Greeting	58.62	61.38
7	Waiting	59.07	79.48
8	WalkTogether	53.13	56.15
9	Phoning	61.27	68.99
10	Discussion	61.36	71.88
11	SittingDown	84.68	**79.43**
12	Eating	56.34	**52.07**
13	Smoking	60.45	**60.25**
14	Directions	54.51	70.50
15	Photo	68.57	82.64
	Avg	*61.8*	*67.08*

According to the result, we can primarily decide the fact that

Table 2. Velocity Error (MPJPE), 3*3, unit-epoch

No.	Action	Original	Action-based
1	Purchases	3.93	**3.65**
2	Posing	3.44	3.79
3	Walking	4.51	**3.77**
4	Sitting	3.14	**2.84**
5	Walkdog	4.68	**4.37**
6	Greeting	4.42	**4.27**
7	Waiting	3.34	**3.55**
8	WalkTogether	4.08	**3.49**
9	Phoning	3.31	**3.14**
10	Discussion	4.02	4.28
11	SittingDown	4.21	**3.90**
12	Eating	3.14	**3.01**
13	Smoking	3.36	**3.05**
14	Directions	3.84	**3.78**
15	Photo	3.58	**3.50**
	Avg	*3.8*	***3.63***

4.2. Large Number of Epochs

For a large number of data, we observe that because of the characteristics of the model that *VideoPose3D* utilizes, the epoch refuses to increase after epoch 60 (i.e. epoch 4 of the original training method).

Table 3. Property Cluster, 3*3, 80-epoch

Type	Original	Eating-based
Time Consumed	? sec	56921 sec
MPJPE-Eating	? mm	49.27 mm
Velo-M-Eating	? mm	2.56 mm

Shown below is the relationship between our work (right) and the original work (left) using protocol MPJPE with the lapse of epochs. Note that each epoch worths $\frac{1}{15}$ Epoch Units.

Table 4. Epoch-based Observations

Epochs	Epoch Units	Original	Eating-based
1	NaN	NaN	138.86
5	NaN	NaN	96.52
10	NaN	NaN	65.02
15	1	?	60.21
30	2	?	52.06
45	3	?	50.28
60	4	?	50.58
75	5	?	50.49
90	6	?	50.37
450	10	?	50.53
1200	80	?	50.45

Transforming data into more readable format we gain the graph below.

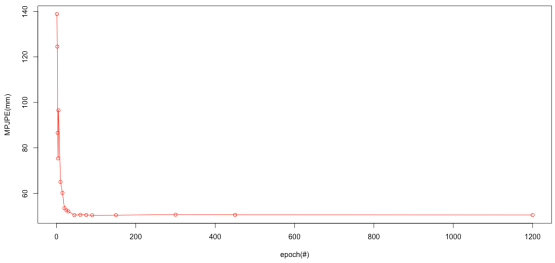

Figure 4. Relationship between MPJPE and epoch (original)

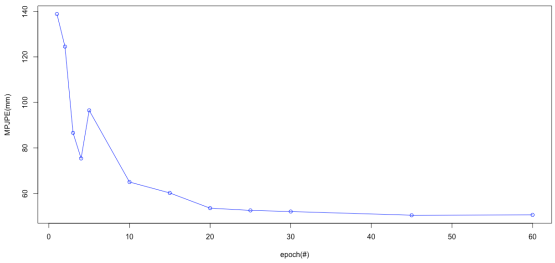

Figure 5. Relationship between MPJPE and epoch (Zoomed)

5. Conclusion

According to the results shown above, we conclude that the action-based model has a better performance on small epochs because they learn faster than the original one, though it has a worse performance on the top performance.

References

[1] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Deep image homography estimation. *arXiv preprint arXiv:1606.03798*, 2016.

[2] Kyoungoh Lee, Inwoong Lee, and Sanghoon Lee. Propagating lstm: 3d pose estimation based on joint interdependency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 119–135, 2018.

[3] Bruce Xiaohan Nie, Caiming Xiong, and Song-Chun Zhu. Joint action recognition and pose estimation from video. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1293–1301, 2015.

[4] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7025–7034, 2017.

[5] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7753–7762, 2019.

[6] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013.

[7] Angela Yao, Juergen Gall, and Luc Van Gool. Coupled action recognition and pose estimation from multiple views. *International journal of computer vision*, 100(1):16–37, 2012.

[8] Xiaojin Zhu and Andrew B Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130, 2009.